Fusion of BLAST and Ensemble of Classifiers for Protein Secondary Structure Prediction

Gabriel Bianchin de Oliveira*, Helio Pedrini*, Zanoni Dias*
* Institute of Computing, University of Campinas, Campinas, SP, Brazil, 13083-852
gabriel.bianchin@students.ic.unicamp.br, helio@ic.unicamp.br, zanoni@ic.unicamp.br

***Abstract*—The prediction of protein secondary structure has great relevance in the analysis of global protein folding. In this work, we present a method for protein secondary structure prediction using the fusion of BLAST and the ensemble of local and global classifiers. We used the amino acid sequence and sequence similarity information available in the datasets and we explored other amino acid characteristics. In order to evaluate our method, we used the files from PDB (only from the year 2018), as well as CB6133 and CB513 datasets. We achieved 87.7%, 82.4% and 85.6% Q8 accuracy on PDB 2018, CB6133 and CB513 proteins using the amino acid sequence and amino acid biological properties, 84.7% and 87.5% Q8 accuracy on CB6133 and CB513 proteins using the amino acid sequence and similarity sequence information and 92.5% Q3 accuracy on PDB 2018 proteins using the amino acid sequence and amino acid biological properties. Our method presented competitive results using only BLAST and only the ensemble of classifiers. The fusion of both approaches achieved superior results compared to state-of-the-art approaches.**

I. Introduction

Proteins are present in several biological processes of living organisms. They are formed by a sequence of amino acids, which, due to the physical and chemical interactions of attraction and repulsion between them, form three-dimensional (3D) structures [1]. The local structure that each amino acid form is called secondary structure.

Secondary structures can be divided into Q3 classification and Q8 classification. In the Q3 classification, each amino acid can be transformed into helix (H), strand (E) or coil (C). With the high accuracy achieved in Q3 classification, the Q8 classification was created. In the Q8 classification, each amino acid can be transformed into 4-turn helix (H), 3-turn helix (G), residue in isolated beta bridge (B), extended strand (E), 5-turn helix (I), hydrogen bonded turn (T), bend (S) and loop (L). The Q8 classification is more complex and challenging than the Q3 classification [2].

From the analysis of the secondary structures of proteins, it is possible to analyze the global 3D structure and the folding of the proteins. With this, it is viable to understand and to create possible applications, such as drug and biosensor design [3], [4].

Due to advances in gene sequencing, there are large volumes of data on the amino acid sequences that make up proteins, but determining 3D structures, such as secondary structures, requires a lot of effort, such as laboratory methods [5]. The difference in the amount of data on the protein sequence and secondary structures can be seen in the data volume of UniProtKB, which is the main protein sequence database and has 175 million data from protein sequences, and Protein Data Bank, which is the main database of secondary protein structures and has 160,000 data from protein secondary structure.

Due to the high cost to determine the secondary protein structures through laboratory methods, other methods have been proposed for the prediction of these structures, with an emphasis on computational approaches [6].

At the beginning of the interest in predicting secondary structures, the main methods used statistical concepts, such as a set of rules [7] and statistical procedures [8], [9].

In the second phase of secondary structure prediction methods, classifiers such as Support Vector Machines (SVM) [10], [11], [12] and Neural Networks (NN) [13], [14] with sliding window achieved rates close to 80% of Q3 accuracy. At this stage, other characteristics for classification began to be explored, for instance, sequence similarity information [15] and amino acid properties [16]. It was also at this stage that the methods began to use the Q8 classification.

The third phase of the classification of secondary structures gained space with the advance of deep learning, mainly with recurrent networks [17], [18], capable of making the global analysis of protein sequences, and convolutional networks [19], [20], generally used for the local analysis of the sequences. Other methods achieved improvements when using global analysis with local analysis [2], [21]. Still in the third phase, several methods in the literature began to study the effect and the improvement of the results through the ensemble of several classifiers [22], [23].

In the literature, BLAST [24], used to align local sequences of proteins, is not commonly used to predict protein secondary structures. In this work, we present and discuss a method for predicting protein secondary structures using a fusion of BLAST with an ensemble of global and local classifiers.

Our main contribution is to present a method that can predict protein secondary structure using amino acid properties, sequence and similarity information. BLAST has good results in the secondary structure prediction, but it cannot predict the secondary structure for all the amino acids, the ensemble classifier achieves competitive results compared to the state of the art, whereas the fusion of BLAST and the ensemble of classifiers can reach superior results.

The paper is organized as follows. We describe our secondary structure prediction method in Section II. The datasets,

amino acid additional properties, evaluation metrics and experimental results are shown in Section III. Concluding remarks and directions for future work are presented in Section IV.

II. Protein Secondary Structure Prediction Method

In this section, we describe our method for protein secondary structure prediction. We divided the method into two parts, BLAST and a classifier, which we called an ensemble of classifiers. The fusion was performed using the bag of optimizers.

A. BLAST

BLAST [24] is a tool that compares the amino acid chain of proteins and finds the best local alignment. To perform a BLAST search, it is necessary to have a query, that is, a protein that serves as the basis for the search, and the database for searching. BLAST can be seen as an information retrieval tool.

To generate the alignment, three different cases can occur: (i) match, that is, the stretch in which there was an alignment, (ii) mismatch, when the alignment does not occur and (iii) gaps, when it is more advantageous not to compare the stretch. Figure 1 illustrates an example of match, mismatch and gap.

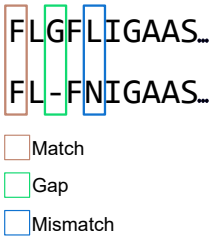

Fig. 1. Example of match, mismatch and gap.

The weight of the matches, mismatches and gaps is given by a substitution matrix. We used the default substitution matrix, BLOSUM62 [25]. In the end, the best alignments receive the highest bit score and the lowest E-value.

We chose to use BLAST to find similar amino acid sequences because proteins with similar sequences can have similar secondary structures. So, we utilized the correspondent secondary structure from the aligned sequence to predict the secondary structure from the query.

In order to search for local sequence of proteins, we applied the PDB as a database, since this dataset is the main benchmark with information on secondary structures. To ensure that the protein found in the PDB is not exactly the same as the search protein, we removed proteins that were the same size and that had the same amino acid sequence as the search protein.

In some alignments, gaps may occur. In these cases, if there was a match between the amino acid and a gap, the structure that corresponded to the gap was disregarded.

We tested various configurations in the search for local protein sequence alignment, such as the selection of the top 5, 10 or 20 alignment of sequences, restriction by E-value and different voting weights for each alignment. Empirically, the best configuration found was to use the top 10 alignment of sequences, employing increasing voting weights for the best alignments and E-value restriction equal to 0.00001.

Using BLAST to find proteins with similar sequence alignment, some amino acids may not have a predicted secondary structure. With that, we employed the fusion with the ensemble of classifiers to predict all cases. If the amino acid did not have a secondary structure predicted by BLAST, the probability vector of each class was equal to 0, else, the probability vector sum was equal to 1. The fusion was performed using the bag of optimizers.

B. Ensemble of Classifiers

The ensemble of classifiers is divided into three parts: local classification, global classification and ensemble of local and global classification. The ensemble was made using the bag of optimizers.

1) Local Classification: In the local classification, we divided the proteins into blocks. Each block contained the central amino acid and the same number of amino acids on the right and on the left. We tested different block sizes and found empirically the best results in block size equal to 3, that is, a central amino acid with an amino acid on the right and an amino acid on the left, up to 11.

At the beginning and at the end of proteins, we explored two different types of padding, values equal to 0 and the repetition of the first element of the sequence at the beginning of the protein and the last element at the end of the protein. Empirically, we obtained better results with the first option in the Q8 classification and with the second option in the Q3 classification.

For classification, five random forests with different block sizes were used. Each of the random forests used a sliding window to traverse each block in the sequence. Then, we merged the five classifiers using weights for each of the classes in each of the random forests. The final prediction of the set of local classifiers was normalized, that is, the sum of the probabilities of all classes was equal to 1. Figure 2 illustrates the methodology proposed for the local classification.

We chose the random forest as a classifier since it improved the fusion with the global classifiers better than the other classifiers, such as convolutional networks.

2) Global Classification: In the global classification, we used bidirectional recurrent networks with GRU memory modules. We chose bidirectional recurrent networks because this configuration can deal with anterior and posterior amino acids in relation to the analyzed amino acid.

For the amino acid sequence features, we applied an embedding layer at the beginning of the network, transforming the sparse vector in the one-hot encoding format into a dense vector. We tested several dense vector sizes and the best

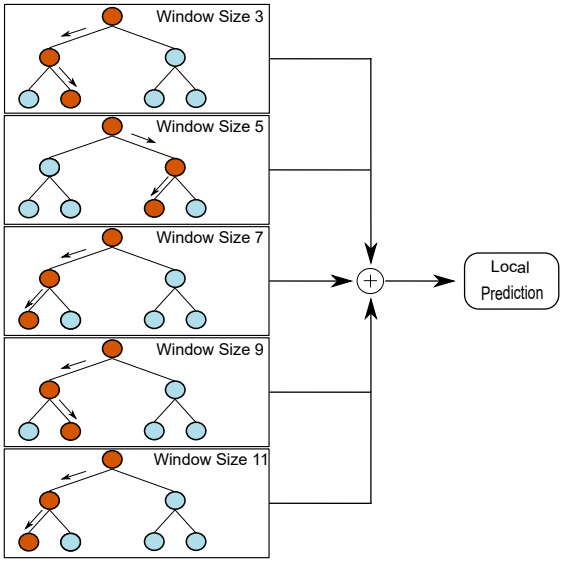

Fig. 2. Local classification method.

configuration was using the dense vector with the same size of the one-hot encoding.

Regarding the number of bidirectional recurrent layers, we explored several configurations and obtained empirically better results with networks with 2 up to 6 recurrent layers for both Q3 and Q8 classification. In the end, the network has a dense layer with softmax activation. We employed Adam [26] optimizer, early-stopping and dropout [27] regularization techniques. Figure 3 illustrates the proposed recurrent network architecture with 2 layers.

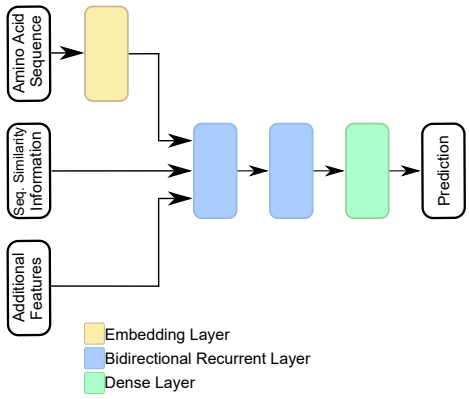

Fig. 3. Architecture of the recurrent network with 2 layers.

For each configuration of the bidirectional recurrent network, we utilized two identical networks, one network analyzing the protein in the standard direction, that is, analyzing the protein from the beginning to the end, and another network analyzing the protein in the reverse direction, that is, analyzing the protein from the end to the beginning. Finally, the prediction of the two networks were concatenated and normalized, so the sum of the probabilities is equal to 1. This methodology proved to be capable of improving the results.

As we did in the local classification, we used weights for each class of each of the networks. Then, the prediction of the set of bidirectional recurrent networks was normalized. Figure 4 shows the methodology proposed for the global classification.

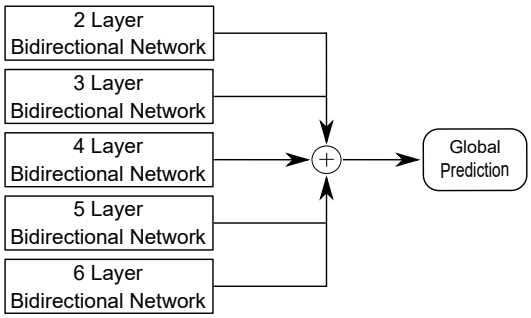

Fig. 4. Global classification method.

3) Ensemble of Local and Global Classification: After the local classification and the global classification, we performed the ensemble of the two classifications. To perform the fusion, each of the classification sets had weights for each of the classes. Finally, the prediction of the ensemble of local and global classifiers was normalized.

C. Bag of Optimizers

For the local and global classification, the ensemble of local and global classification and the fusion of BLAST prediction with the ensemble of classifiers prediction, we applied the bag of optimizers with three different algorithms. Then, the best weight found among them was chosen. The bag of optimizers was used as a black box.

The first algorithm in the bag of optimizers was the genetic algorithm. Initially, we created a population of size equal to 2,000 with weights ranging from 0 to 10. The top 100 individuals become parents of the next generation. The parents generated 900 new individuals from crossover, totaling the first 1,000 individuals of the next generation. The 1,000 individuals generated other 1,000 individuals through mutation. This process was carried out for 100 generations. At the end of 100 generations, the top 100 individuals generated 900 new individuals through mutation. This process was carried out for 100 generations. In the end, the best individual was chosen.

The second algorithm in the bag of optimizers was the cuckoo search [28]. Initially we created a population of size equal to 1,000 with weights ranging from 0 and 10. For each individual, we obtained the corresponding cuckoo using the levy flights. For each cuckoo, we checked another random cuckoo and performed the replacement if the random cuckoo found was worse than the cuckoo that performed the search. This process occurred 100 times. With each iteration, the 250 worst cuckoos were reset. In the end, the best cuckoo was chosen.

The third algorithm in the bag of optimizers was the particle swarm optimization [29]. Initially, we created a population of

size equal to 1,000 with weights ranging from 0 to 10. For each individual, we calculated the best result obtained. Then, we calculated the best overall result obtained by any individual. In the end, we updated the individuals taking into account the best personal result, the best global result and the direction in which the individual was moving. This process occurred 100 times. In the end, the best individual was chosen.

III. Experiments

In this section, we present the datasets and the amino acid properties used in our experiments, as well as the evaluation metrics and the results obtained.

A. Datasets

In this subsection, we describe the datasets used in the experiments.

1) PDB: The Protein Data Bank, known as PDB, is the main repository for 3D structures of proteins. This repository has more than 150,000 structures of proteins, nucleic acids and complex macromolecules. It receives weekly updates.

From the PDB, we selected proteins up to 700 amino acids from the year 2018 (called PDB 2018 from now on). We applied the same split as proposed by Oliveira et al. [30], that is, [0, 6478] proteins for training, [6479, 6978] proteins for validation and [6979, 7478] proteins for testing.

Some amino acids have two letters to represent them. Therefore, we considered the amino acid "X" as the amino acid "A", the amino acid "B" as the amino acid "N" and the amino acid "Z" as the amino acid "Q".

To generate the secondary protein structures, we used the DSSP tool [31], [32]. In the Q3 classification, we considered the G and H classes from Q8 classification as H, B and E classes from Q8 classification as E and I, L, S, T classes from Q8 as C. The classes in Q3 and Q8 classification are unbalanced.

2) CB6133: The CB6133 database is a set of 6,133 proteins of 50 to 700 amino acids, available on the PISCES CullPDB [33] server. The proteins that are part of the set have less than 30% similarity between them [3].

In this database, the amino acid "X" is different from the amino acid "A". Therefore, it has 21 amino acid sequence information in the one-hot encoding format. In addition to the amino acid sequences, the CB6133 dataset has similarity sequence information, which was generated using the PSI-BLAST [24] against the UniRef90 database with a 0.001 threshold and 3 iterations. To transform the similarity data between 0 and 1, the sigmoid function was applied [3]. Similarity sequence information cannot be produced for large databases in a timely manner [34].

We employed the same split used in the literature, that is, [0, 5599] proteins for training, [5877, 6132] proteins for validation and [5605, 5876] for testing. Classes are unbalanced and there is no "I" structure in the test set.

A filtered version of this dataset was utilized to train and validate the testing on CB513. This filtered version has proteins with less than 25% of similarity to CB513 database. We split the filtered version into [0, 5277] proteins for training and [5278, 5533] proteins for validation.

3) CB513: The CB513 dataset [35] has 513 proteins. In this dataset, a protein has more than 700 amino acids, so we truncated this protein to 700 amino acids and the remaining was considered another protein.

We utilized this database for prediction (testing). For training, we employed the filtered version of CB6133. The CB513 dataset has 21 features from amino acid sequence and 21 features from sequence similarity information. Classes are unbalanced.

B. Amino Acid Properties

In this subsection, we present the additional amino acid properties used in the protein secondary structure prediction.

1) Amino Acid Biological Properties: As additional characteristics for classification, we used 8 different biological properties of amino acids, as employed by Pok et al. [16].

For each amino acid in the sequence, we assigned a feature vector represented whether or not the amino acid has the specific feature. We applied the characteristics in relation to hydropathy, charged or uncharged, size and polar or non-polar. Values equal to 1 indicated the presence of the characteristic and values equal to 0 indicated the lack of the characteristic. In the case of the amino acid "X", we considered it the same as the amino acid "A".

In the padding process, we assigned all 8 different biological properties with values equal to 0.

2) Distance: For each of the amino acids in the protein sequence, we calculated the shortest distance from the analyzed amino acid with all other different amino acids, checking the anterior and posterior amino acids. If there was no specific amino acid in the sequence, the distance to it was infinite.

With the distances calculated for all amino acids, we normalized the values by applying the hyperbolic tangent function. In the padding, we set all features with values equal to 0.

3) Statistical Measures: In order to create statistical measures as additional characteristics, we used window of size 11, that is, five anterior amino acids, the analyzed amino acid and five subsequent amino acids, 21 and 41. For amino acids at the beginning or at the end of the protein, we considered only the nearby amino acids and excluded the padding.

Within each window, we calculated the mean, mode and median, normalizing the values for the interval between 0 and 1. In the padding, we set all features with values equal to 0.

C. Evaluation Metrics

The performance of the proposed method was assessed through the following evaluation metrics. In their formulation, TP is the number of true positive cases, FP is the number of false positive cases and FN is the number of false negative cases.

Equation (1) presents the precision metric. We used this metric to evaluate each class.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{1}$$

Equation (2) presents the recall metric. We applied this metric to evaluate each class.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (2)$$

Equation (3) presents the Q3 accuracy metric. We employed this metric to evaluate the method in the Q3 classification.

$$\text{Accuracy}_{\text{Q3}} = \frac{\sum_{i \epsilon \{\text{Q3 classes}\}} \text{correct predictions in } i}{\sum_{i \epsilon \{\text{Q3 classes}\}} \text{residues in } i} \quad (3)$$

Equation (4) presents the Q8 accuracy metric. We utilized this metric to evaluate the method in the Q8 classification.

$$\text{Accuracy}_{\text{Q8}} = \frac{\sum_{i \epsilon \{\text{Q8 classes}\}} \text{correct predictions in } i}{\sum_{i \epsilon \{\text{Q8 classes}\}} \text{residues in } i} \quad (4)$$

D. Results

We divided the experiments into three parts: (i) training and testing on PDB, (ii) training and testing on CB6133 and (iii) training on filtered CB6133 and testing on CB513.

1) Training and Testing on PDB: In the Q3 classification and the Q8 classification, we employed BLAST with the top 10 alignments, using increasing voting weights for the best alignments and E-value restriction equal to 0.00001. Since the proteins are the same on Q3 and Q8 classification, BLAST predicted 122,243 secondary structures from the test set (84.5%), which has 500 proteins and 144,682 amino acids. The Q3 accuracy was 93.7% and the Q8 accuracy was 89.3%.

After, we utilized the ensemble of classifiers. In order to make a fair comparison with the literature, first we employed only the amino acid sequence. For the Q3 classification, the ensemble of classifiers achieved 82.6% of Q3 accuracy using 800 neurons in the recurrent network. The best result in the literature for PDB 2018 proteins was 81.5% [30]. Table I reports the precision and recall rates for each class.

TABLE I
PRECISION AND RECALL RATES ON PDB 2018 FOR EACH CLASS IN THE Q3 AND Q8 CLASSIFICATION USING THE ENSEMBLE OF CLASSIFIERS (EC) AND THE STATE-OF-THE-ART APPROACH [30].

Classification	Class	Precision		Recall	
		[30]	EC	[30]	EC
Q3	C	0.79	**0.81**	**0.84**	0.83
	E	**0.81**	0.80	0.69	**0.75**
	H	0.85	**0.86**	0.86	**0.87**
Q8	B	0.82	0.82	**0.44**	0.41
	E	0.72	**0.74**	0.79	**0.80**
	G	**0.81**	0.79	0.54	**0.55**
	H	0.78	**0.81**	0.91	**0.92**
	I	**0.85**	0.81	**0.57**	0.56
	L	0.66	**0.67**	0.70	**0.71**
	S	**0.79**	0.76	0.39	**0.40**
	T	**0.64**	0.63	0.55	**0.58**

For the Q8 classification, the ensemble of classifiers achieved 74.3% of Q8 accuracy using 900 neurons in the recurrent network. The best result in the literature to PDB 2018 proteins was 73.1% [30]. Table I reports the precision and recall rates for each class.

After, we utilized the additional features for the classification. We used different configurations of the additional features and the best result was 82.8% of Q3 accuracy and 74.3% of Q8 accuracy. In both of them, we employed the amino acid sequence and biological properties.

Then, we applied the fusion of BLAST and the ensemble of classifiers. In the Q3 classification, we used the amino acid sequence and biological properties in the ensemble of classifiers and achieved 92.5% of Q3 accuracy. Table II shows the precision and recall rates for each class in the Q3 classification of the fusion (BLAST+EC) and the state-of-the-art [30] approach.

TABLE II
PRECISION AND RECALL RATES ON PDB 2018 FOR EACH CLASS IN THE Q3 AND Q8 CLASSIFICATION USING BLAST AND ENSEMBLE OF CLASSIFIERS (BLAST+EC) AND THE STATE-OF-THE-ART APPROACH [30].

Classification	Class	Precision		Recall	
		[30]	BLAST+EC	[30]	BLAST+EC
Q3	C	0.79	**0.92**	0.84	**0.91**
	E	0.81	**0.92**	0.69	**0.92**
	H	0.85	**0.94**	0.86	**0.95**
Q8	B	**0.82**	0.78	0.44	**0.71**
	E	0.72	**0.91**	0.79	**0.92**
	G	**0.81**	0.80	0.54	**0.77**
	H	0.78	**0.93**	0.91	**0.96**
	I	**0.85**	0.82	0.57	**0.78**
	L	0.66	**0.86**	0.70	**0.86**
	S	0.79	0.79	0.39	**0.72**
	T	0.64	**0.78**	0.55	**0.78**

In the Q8 classification, we utilized BLAST and the ensemble of classifiers with the amino acid sequence and biological properties and achieved 87.7% of Q8 accuracy. Table II presents the precision and recall rates for each class in the Q8 classification of the fusion (BLAST+EC) and the state-of-the-art [30] approach.

2) Training and Testing on CB6133: In this experiment, we performed the Q8 classification with the methods using BLAST, the ensemble of classifiers and the fusion of both.

Initially, we used BLAST in order to find the similar alignments of proteins of the test set of CB6133 and the proteins from the PDB dataset. In the test set, there were 272 proteins and 56,686 amino acids and secondary structures. With BLAST using the top 10 alignments, using increasing voting weights for the best alignments and E-value restriction equal to 0.00001, 45,562 secondary structures were predicted (80.4%), with a Q8 accuracy equal to 86.5%.

Then, we applied the ensemble of classifiers. First, we utilized only the features in the dataset in order to make a fair comparison with the literature. The weights of the ensemble were found in the validation set. Using only the amino acid

sequence, the ensemble of classifiers achieved 61.6% of Q8 accuracy with 900 neurons in each layer of the recurrent neural network, surpassing the best result in the literature using only amino acid sequence (59.1%) [30].

Then, we used the amino acid sequence and sequence similarity information in the ensemble of classifiers. Empirically, we obtained the best result with 900 neurons in each layer of the recurrent neural network. We obtained 75.8% of Q8 accuracy. Table III presents our result compared to other results available in the literature. Table IV shows the precision and recall rates for our method and state-of-the-art [22] approach.

TABLE III
Q8 ACCURACY ON CB6133 DATASET.

Methods	Q8 Accuracy (%)
BLAST+EC (Fusion 2)	**84.7**
BLAST+EC (Fusion 1)	**82.4**
Ensemble of Methods [22]	76.3
Ensemble of Classifiers (EC)	**75.8**
2DConv-BLSTM [6]	75.7
biRNN-CRF [36]	74.8
DeepACLSTM [2]	74.2
CNNH_PSS [37]	74.0
Ensemble of RNN and RF [30]	73.4
GSN [3]	72.1

TABLE IV
PRECISION AND RECALL RATES ON CB6133 DATASET FOR EACH CLASS IN THE Q8 CLASSIFICATION USING THE ENSEMBLE OF CLASSIFIERS (EC) AND THE STATE-OF-THE-ART APPROACH [22].

Class	Precision		Recall	
	[22]	EC	[22]	EC
B	0.66	**0.74**	0.07	**0.20**
E	0.80	**0.81**	0.85	0.85
G	0.54	**0.56**	0.33	**0.35**
H	0.87	0.87	0.94	**0.95**
I	—	—	—	—
L	0.58	**0.63**	0.68	0.68
S	**0.59**	0.54	0.23	**0.34**
T	0.58	**0.63**	0.59	0.59

After, we evaluated the incorporation of other features into the model. Initially, we considered only the amino acid sequence (without sequence similarity information) and we tested several additional characteristics (amino acid biological properties, distance and statistical measures). The best result obtained was using only biological properties, with 900 neurons in each layer of the recurrent network, which achieved 62.0% of Q8 accuracy.

Then, we utilized the amino acid sequence and the sequence similarity information. We tested several additional features (amino acid biological properties, distance and statistical measures), but we did not improve the result (EC) shown in Table III.

Finally, we applied the fusion of the ensemble of classifiers and BLAST. The weights were found in the validation set. Along with BLAST, we used the two best configurations of the ensemble of classifiers, that is, using amino acid sequence and biological properties (Fusion 1) and using amino acid sequence and sequence similarity information (Fusion 2). We achieved 82.4% of Q8 accuracy with the Fusion 1 and achieved 84.7% of Q8 accuracy with Fusion 2. Table III shows our result compared to other results available in the literature. Table V presents the precision and recall rates for Fusion 1, Fusion 2 and the state-of-the-art [22] approach.

TABLE V
PRECISION AND RECALL RATES ON CB6133 DATASET FOR EACH CLASS IN THE Q8 CLASSIFICATION USING BLAST AND ENSEMBLE OF CLASSIFIERS (FUSION 1 AND FUSION 2) AND THE STATE-OF-THE-ART APPROACH [22].

Class	Precision			Recall		
	[22]	Fusion 1	Fusion 2	[22]	Fusion 1	Fusion 2
B	0.66	0.67	**0.71**	0.07	0.63	0.63
E	0.80	0.87	**0.91**	0.85	0.90	**0.91**
G	0.54	0.71	**0.72**	0.33	0.59	**0.63**
H	0.87	0.90	**0.93**	0.94	0.93	0.94
I	—	—	—	—	—	—
L	0.58	0.77	0.77	0.68	0.76	**0.80**
S	0.59	0.68	**0.71**	0.23	0.62	**0.63**
T	0.58	0.72	**0.74**	0.59	0.70	**0.73**

3) Training on Filtered CB6133 and Testing on CB513: Initially, we used BLAST to find the similar protein alignments from the CB513 test set and PDB proteins. In the test set, there were 514 proteins and 84,765 amino acids and secondary structures. With BLAST using the top 10 alignments, using increasing voting weights for the best alignments and E-value restriction equal to 0.00001, 72,341 secondary structures were predicted (85.3%), with Q8 accuracy equal to 90.2%.

Then, we applied the ensemble of classifiers, using the filtered version of CB6133 for training and validate and CB513 for testing. As we did in the training and testing on CB6133, first we made a comparison with the literature using only the features of the dataset.

First, we utilized only the amino acid sequence. The weights of the ensemble were found in the validation set. Empirically, we obtained the best result with 500 neurons in each layer of the recurrent neural network and we achieved 57.4% of Q8 accuracy. The best result reported in the literature using only amino acid sequence was 57.1% [2].

After, we employed the amino acid sequence and the sequence similarity information. The weights of the ensemble were found in the validation set. We obtained the best result with 900 neurons in the layers of the recurrent bidirectional network. We obtained 71.2% of Q8 accuracy. Table VI presents our result against the results available in the literature. Table VII shows the precision and recall rates for our method and state-of-the-art [38] approach.

Then, we tested the incorporation of additional features in the model. First, we considered only the amino acid sequence (without sequence similarity information) and we evaluated several additional features (amino acid biological properties, distance and statistical measures). We obtained the best result using only the biological properties, with 500 neurons in the recurrent network. We achieved 57.6% of Q8 accuracy.

TABLE VI
Q8 ACCURACY ON CB513 DATASET.

Methods	Q8 Accuracy (%)
BLAST+EC (Fusion 2)	**87.5**
BLAST+EC (Fusion 1)	**85.6**
Conditioned CNN [38]	71.4
Ensemble of Classifiers (EC)	**71.2**
DeepNRN [39]	71.1
biRNN-CRF [36]	70.9
Ensemble of Methods [22]	70.9
Ensemble of RNN and RF [30]	68.9
BLSTM [18]	67.4
GSN [3]	66.4
CNF [40]	63.3
BRNN [41]	51.1

TABLE VII
PRECISION AND RECALL RATES ON CB513 DATASET FOR EACH CLASS IN THE Q8 CLASSIFICATION USING THE ENSEMBLE OF CLASSIFIERS (EC) AND THE STATE-OF-THE-ART APPROACH [38].

Class	Precision		Recall	
	[38]	EC	[38]	EC
B	**0.79**	0.61	0.05	**0.09**
E	**0.78**	0.75	0.84	**0.85**
G	**0.53**	0.45	0.29	**0.32**
H	0.85	0.85	**0.94**	0.93
I	0.00	0.00	0.00	0.00
L	0.57	**0.61**	**0.71**	0.65
S	**0.62**	0.55	0.24	**0.29**
T	**0.59**	0.56	0.54	**0.55**

After, we utilized the amino acid sequence and the sequence similarity information. We evaluated several additional features, but we did not improve the result (EC) shown in Table VI.

Finally, we used the fusion of BLAST and the ensemble of classifiers. The weights of the fusion were found in the validation set. Along with BLAST, we used the two best configurations of the ensemble of classifiers, that is, using amino acid sequence and biological properties (Fusion 1) and using amino acid sequence and sequence similarity information (Fusion 2). We achieved 85.6% of Q8 accuracy with the Fusion 1 and we obtained 87.5% of Q8 accuracy with the Fusion 2. Table VI shows our result compared to other results available in the literature. Table VIII presents the precision and recall rates for Fusion 1, Fusion 2 and the state-of-the-art [38] approach.

IV. CONCLUSIONS AND FUTURE WORK

The prediction of secondary protein structures has a major impact on the analysis of protein folding. Even with several methods in the literature, there is no method that solves the problem with great results.

In this work, we presented a fusion of the prediction made by BLAST with the ensemble of classifiers. The predictions obtained with BLAST achieved good results, however, it is not possible to classify all structures. The predictions achieved with the ensemble of classifiers are competitive compared to the state-of-the-art results. With the fusion of BLAST prediction and the ensemble of classifiers predictions, we achieved results that are superior to those available in the literature.

TABLE VIII
PRECISION AND RECALL RATES ON CB513 DATASET FOR EACH CLASS IN THE Q8 CLASSIFICATION USING BLAST AND ENSEMBLE OF CLASSIFIERS (FUSION 1 AND FUSION 2) AND THE STATE-OF-THE-ART APPROACH [38].

Class	Precision			Recall		
	[38]	Fusion 1	Fusion 2	[38]	Fusion 1	Fusion 2
B	**0.79**	0.75	0.77	0.05	**0.71**	0.69
E	0.78	0.88	**0.92**	0.84	0.92	**0.94**
G	0.53	0.74	0.74	0.29	0.72	**0.76**
H	0.85	0.92	**0.94**	0.94	0.95	**0.96**
I	0.00	0.00	0.00	0.00	0.00	0.00
L	0.57	0.83	**0.84**	0.71	0.83	**0.86**
S	0.62	0.79	**0.81**	0.24	0.69	**0.70**
T	0.59	0.77	**0.78**	0.54	0.75	**0.78**

The fusion weights of BLAST and the ensemble of classifiers followed the ratio close to 1 for BLAST and 0.8 for the ensemble of classifiers, showing that both classifiers helped in the final prediction.

The use of sequence alignment proved to be a path for future work, as well as the application of other additional characteristics of the amino acids and the protein sequence.

ACKNOWLEDGMENTS

The authors would like to thank FAPESP (grants #2015/11937-9, #2017/12646-3, #2017/16246-0, #2017/12646-3 and #2019/20875-8), CNPq (grants #304380/2018-0 and #309330/2018-1) and CAPES for their financial support.

REFERENCES

[1] C. Zhou, C. Sun, B. Wang, and X. Wang, "An improved stochastic fractal search algorithm for 3D protein structure prediction," *Journal of Molecular Modeling*, vol. 24, no. 6, p. 125, 2018.

[2] Y. Guo, W. Li, B. Wang, H. Liu, and D. Zhou, "DeepACLSTM: deep asymmetric convolutional long short-term memory neural models for protein secondary structure prediction," *BMC Bioinformatics*, vol. 20, no. 1, pp. 1–12, 2019.

[3] J. Zhou and O. Troyanskaya, "Deep Supervised and Convolutional Generative Stochastic Network for Protein Secondary Structure Prediction," in *31st International Conference on Machine Learning (ICML)*, 2014, pp. 745–753.

[4] H. Kamisetty and C. J. Langmead, "A Bayesian approach to protein model quality assessment," in *26th Annual International Conference on Machine Learning (ICML)*. ACM, 2009, pp. 481–488.

[5] P. Kumar, S. Bankapur, and N. Patil, "An enhanced protein secondary structure prediction using deep learning framework on hybrid profile based features," *Applied Soft Computing*, vol. 86, p. 105926, 2020.

[6] Y. Guo, B. Wang, W. Li, and B. Yang, "Protein secondary structure prediction improved by recurrent neural networks integrated with two-dimensional convolutional neural networks," *Journal of Bioinformatics and Computational Biology*, vol. 16, no. 5, p. 1850021, 2018.

[7] P. Y. Chou and G. D. Fasman, "Prediction of protein conformation," *Biochemistry*, vol. 13, no. 2, pp. 222–245, 1974.

[8] J. Garnier, D. J. Osguthorpe, and B. Robson, "Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins," *Journal of Molecular Biology*, vol. 120, no. 1, pp. 97–120, 1978.

[9] J.-F. Gibrat, J. Garnier, and B. Robson, "Further developments of protein secondary structure prediction using information theory: new parameters and consideration of residue pairs," *Journal of Molecular Biology*, vol. 198, no. 3, pp. 425–443, 1987.

[10] S. Hua and Z. Sun, "A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach," *Journal of Molecular Biology*, vol. 308, no. 2, pp. 397–407, 2001.

[11] W. Zhong, J. He, R. Harrison, P. C. Tai, and Y. Pan, "Clustering support vector machines for protein local structure prediction," *Expert Systems with Applications*, vol. 32, no. 2, pp. 518–526, 2007.

[12] B. Yang, Q. Wu, Z. Ying, and H. Sui, "Predicting protein secondary structure using a mixed-modal SVM method in a compound pyramid model," *Knowledge-Based Systems*, vol. 24, no. 2, pp. 304–313, 2011.

[13] L. H. Holley and M. Karplus, "Protein secondary structure prediction with a neural network," *National Academy of Sciences*, vol. 86, no. 1, pp. 152–156, 1989.

[14] Z. Zhang and N. Jing, "Radial basis function method for prediction of protein secondary structure," in *7th International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 3. IEEE, 2008, pp. 1379–1383.

[15] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *Journal of Molecular Biology*, vol. 292, no. 2, pp. 195–202, 1999.

[16] G. Pok, C. H. Jin, and K. H. Ryu, "Correlation of amino acid physico-chemical properties with protein secondary structure conformation," in *7th International Conference on BioMedical Engineering and Informatics (BMEI)*, vol. 1. IEEE, 2008, pp. 117–121.

[17] L. T. Hattori, C. M. V. Benitez, and H. S. Lopes, "A deep bidirectional long short-term memory approach applied to the protein secondary structure prediction problem," in *IEEE Latin American Conference on Computational Intelligence (LACCI)*. IEEE, 2017, pp. 1–6.

[18] S. K. Sønderby and O. Winther, "Protein secondary structure prediction with long short term memory networks," *arXiv preprint arXiv:1412.7828*, 2014.

[19] C. Fang, Y. Shang, and D. Xu, "MUFOLD-SS: New deep inception-inside-inception networks for protein secondary structure prediction," *Proteins: Structure, Function, and Bioinformatics*, vol. 86, no. 5, pp. 592–598, 2018.

[20] M. R. Uddin, S. Mahbub, M. S. Rahman, and M. S. Bayzid, "SAINT: self-attention augmented inception-inside-inception network improves protein secondary structure prediction," *bioRxiv*, p. 786921, 2019.

[21] M. Torrisi, M. Kaleel, and G. Pollastri, "Deeper profiles and cascaded recurrent and convolutional neural networks for state-of-the-art protein secondary structure prediction," *Scientific Reports*, vol. 9, no. 1, pp. 1–12, 2019.

[22] I. Drori, I. Dwivedi, P. Shrestha, J. Wan, Y. Wang, Y. He, A. Mazza, H. Krogh-Freeman, D. Leggas, and K. Sandridge, "High quality prediction of protein Q8 secondary structure by diverse neural network architectures," *arXiv preprint arXiv:1811.07143*, 2018.

[23] S. Long and P. Tian, "Protein secondary structure prediction with context convolutional neural network," *RSC Advances*, vol. 9, no. 66, pp. 38 391–38 396, 2019.

[24] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[25] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *National Academy of Sciences*, vol. 89, no. 22, pp. 10 915–10 919, 1992.

[26] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference for Learning Representations (ICLR)*, 2015, pp. 254–269.

[27] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[28] X.-S. Yang and S. Deb, "Cuckoo search via Lévy flights," in *1st World Congress on Nature & Biologically Inspired Computing (NaBIC)*. IEEE, 2009, pp. 210–214.

[30] G. B. Oliveira, H. Pedrini, and Z. Dias, "Ensemble of bidirectional recurrent networks and random forests for protein secondary structure prediction," in *27th International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE, 2020, pp. 311–316.

[31] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers: Original Research on Biomolecules*, vol. 22, no. 12, pp. 2577–2637, 1983.

[32] W. G. Touw, C. Baakman, J. Black, T. A. Te Beek, E. Krieger, R. P. Joosten, and G. Vriend, "A series of PDB-related databanks for everyday needs," *Nucleic Acids Research*, vol. 43, no. D1, pp. D364–D368, 2015.

[33] G. Wang and R. L. Dunbrack Jr, "PISCES: A protein sequence culling server," *Bioinformatics*, vol. 19, no. 12, pp. 1589–1591, 2003.

[34] C. Fang, Y. Shang, and D. Xu, "MUFold-SS: Protein secondary structure prediction using deep inception-inside-inception networks," *arXiv preprint arXiv:1709.06165*, 2017.

[35] J. A. Cuff and G. J. Barton, "Evaluation and improvement of multiple sequence methods for protein secondary structure prediction," *Proteins: Structure, Function, and Bioinformatics*, vol. 34, no. 4, pp. 508–519, 1999.

[36] A. R. Johansen, C. K. Sønderby, S. K. Sønderby, and O. Winther, "Deep recurrent conditional random field network for protein secondary prediction," in *8th Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM BCB)*, 2017, pp. 73–78.

[37] J. Zhou, H. Wang, Z. Zhao, R. Xu, and Q. Lu, "CNNH_PSS: Protein 8-class secondary structure prediction by convolutional neural network with highway," *BMC Bioinformatics*, vol. 19, no. 4, p. 60, 2018.

[38] A. Busia and N. Jaitly, "Next-step conditioned deep convolutional neural networks improve protein secondary structure prediction," *arXiv preprint arXiv:1702.03865*, 2017.

[39] C. Fang, Y. Shang, and D. Xu, "A new deep neighbor residual network for protein secondary structure prediction," in *IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2017, pp. 66–71.

[40] Z. Wang, F. Zhao, J. Peng, and J. Xu, "Protein 8-class secondary structure prediction using conditional neural fields," in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2010, pp. 109–114.

[41] G. Pollastri, D. Przybylski, B. Rost, and P. Baldi, "Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles," *Proteins: Structure, Function, and Bioinformatics*, vol. 47, no. 2, pp. 228–235, 2002.